

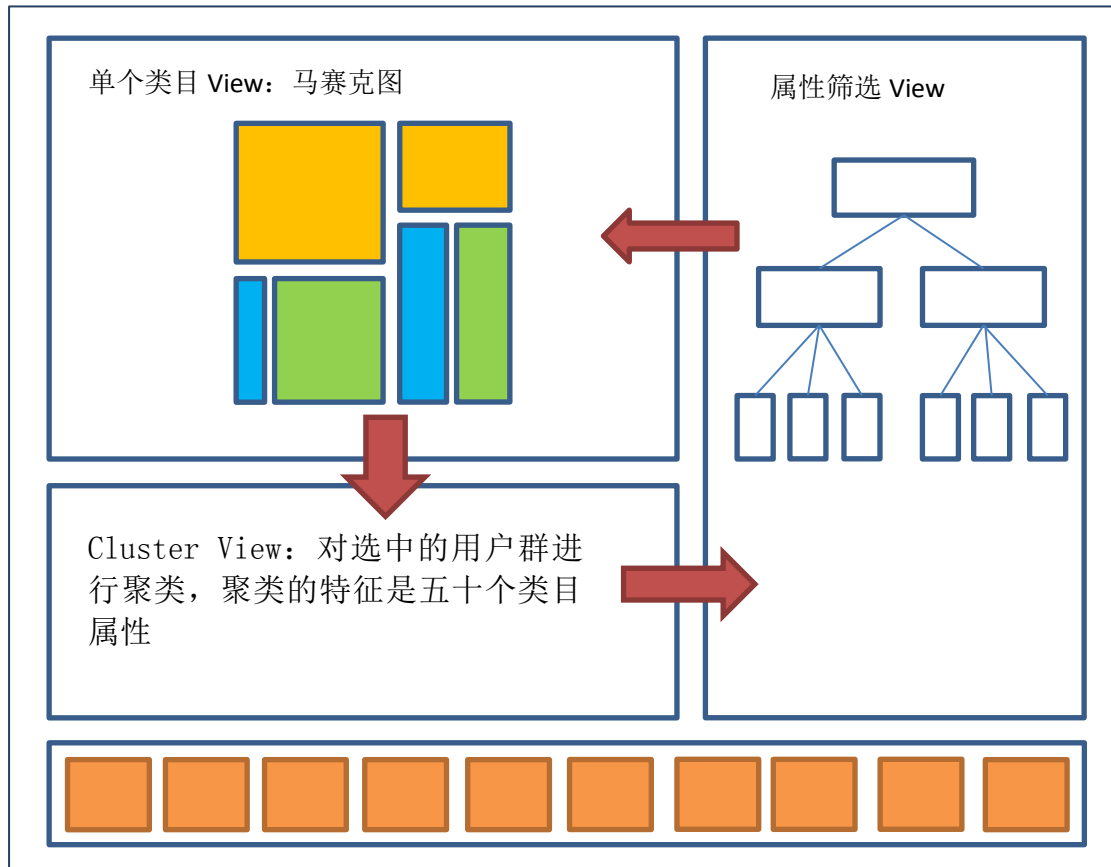
Weekly Report

2012.10.8-2012.10.14

黄芯芯

本周工作：

1. 周三的论文报告。
2. 课程作业：学习 GlusterFS 分布式文件系统。
3. 淘宝标签项目新方向：
 - 1) 框架如下图所示：



属性筛选器可以通过两种方式对属性进行筛选，系统推荐或者用户自定义；根据筛选的分类在每个类目上对全部用户进行一个马赛克图显示；用户可对马赛克图进行交互，选中部分人群，然后对这些人群进行再一次聚类分析，分析得到的结果可能会存在异常值或者其他情况，用户再通过交互将某部分结果回送到筛选器中，从而形成一个 Loop。

对于筛选之后的分类的用户群，每一类可以当成是一个文档，全部用户就是整个文本集，因此可以尝试文本主题挖掘的方法，找出最能代表每类用户的类目特征。和丙辉了解了 LDA，LDA 需要非常大量的输入文本数据，如果数据量不够大可能挖掘不出有用的主题，而且学习时间比较长。所以，对于使用文本主题抽取这一块可能用最简单的 TFIDF 方法尝试。

关于聚类，从屈老师 DICON 文章中得到的启发是，聚类之后要把每个聚类的特征显示出来，让用户可以理解该聚类结果的语义。

- 2) 准备用 java+prefuse 实现系统，因此这几天也在学习 prefuse。

下周计划:

1. 继续学习 `prefuse`，开始实现系统模块。
2. 上课&课程作业。